整体报告

📚 Step 1: Exploratory Data Analysis (EDA)

1.0 导入（Libraries Import）

- 引入常用的数据处理、绘图和机器学习库。
- 提前导入模型相关库，不影响EDA后续流程。
- 主要使用 `numpy`、`pandas`、`matplotlib`、`seaborn`、`scikit-learn`。

1.1 加载数据（Load Data）

- 使用 `pd.read_csv` 读取原始数据集 `custom_covid19.csv`。
- 打印数据集大小（shape）和前5行样本。
- 使用 `df.info()` 查看每列数据类型和非空数量，初步了解数据结构。

1.2 识别变量类型（Variable Types）

- 粗略将特征分为：
  - `categorical_features`：分类特征。
  - `numerical_features`：数值特征（本数据集中仅 `AGE`）。
  - `special_features`：需要特别处理的特征（如 `DATE_DIED`）。
- 生成新变量 `DIED`：
  - 如果 `DATE_DIED == '9999-99-99'`，则标记为 0（未死亡），否则标记为 1（已死亡）。

1.3 缺失值检测（Missing Values Detection）

- 将特殊编码（97, 98, 99）识别为缺失。
- 针对 `DATE_DIED`，将 `'9999-99-99'` 统计为缺失。
- 可选绘制缺失值数量的条形图。

1.4 异常值检测（Outlier Detection）

- 使用 IQR 方法（四分位数）检测异常值（仅针对 `AGE`）。
- 打印异常值数量和边界。
- 使用 `seaborn` 画箱型图（Boxplot）可视化异常值分布。
- **小结**：`AGE` 有部分异常值，但在项目中未直接剔除，后续模型可自动处理。

1.5 单变量分布（Variable Distributions）

- 绘制关键变量（如 `AGE`，`SEX`，`PATIENT_TYPE`，`ICU` 等）分布图。
- 对于分类变量使用 `countplot`，对连续变量使用 `histplot`。
- 帮助直观了解变量取值特性。

1.6 特征间相关性分析（Correlation Analysis）

- 选取数值型特征，计算相关系数矩阵（Pearson相关性）。
- 使用 `seaborn.heatmap` 绘制热力图，观察特征之间的线性关系。
- 特别查看：
  - **AGE** 和其他特征的相关性（用于 O2）。
  - **TEST_RESULT** 和其他特征的相关性（用于 O3）。

1.7 重要特征标记（Important Feature Recognition）

- 根据医学常识和相关性，人工初步挑选重要特征：

任务	重要特征
O1（死亡预测）	AGE, INTUBED, ICU, PNEUMONIA, DIABETES, COPD, HYPERTENSION, OBESITY, RENAL_CHRONIC, CARDIOVASCULAR
O2（年龄预测）	DIABETES, COPD, TOBACCO, HYPERTENSION, CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, ASTHMA, INMSUPR
O3（死亡患者年龄预测）	ICU, INTUBED, PNEUMONIA, DIABETES, HYPERTENSION, OBESITY, RENAL_CHRONIC, COPD, TOBACCO, ASTHMA, INMSUPR

1.8 分组特征均值对比（Feature Group Means）

- 对重要特征按 `DIED` 分组，计算均值。
- 绘制每个特征在生存者与死亡者中的均值柱状图。
- **观察**：
  - 死亡者在 `AGE`、`ICU`、`INTUBED`、`PNEUMONIA` 等特征上的均值明显更高。

1.9 简易Baseline决策树（Baseline Decision Tree）

- 使用初步筛选的7个特征训练简单决策树（`max_depth=3`）。
- 评估验证集上的性能，查看模型基础效果。
- 主要作为后续更复杂建模的对比参考。

1.10 缺失值处理（Missing Value Handling）

- 全面替换 97, 98, 99 为 `np.nan`。
- 分类特征用众数（mode）填充。
- 数值特征用中位数（median）填充。
- 保证后续建模过程中不会因缺失值出错。

这里的填充是很粗糙的填充

1.11 数据格式修正并保存（Fix Data Types & Save Cleaned Dataset）

- 将应为整数但被误读为浮点型的列转为 `Int64` 类型。
- 统一数据格式后，保存干净版数据集：
  `custom_covid19_cleaned.csv`

Step 2: Task O1 - Death Prediction (Classification)

2.1 数据准备（Data Preparation）

- 选择特征集 `important_features_O1`：

```
["AGE", "INTUBED", "ICU", "PNEUMONIA", "DIABETES", "COPD",
 "HYPERTENSION", "OBESITY", "RENAL_CHRONIC", "CARDIOVASCULAR"]
```

- 目标列：`DIED`
- 替换缺失值（97, 98, 99）为众数。
- 划分训练集（80%）与验证集（20%），保证分布一致（stratify）。

2.2 模型训练与初步评估（Baseline Models）

- 训练模型一：**Decision Tree Classifier**
  - 限制最大深度为5，防止过拟合。
  - 评估验证集，得到基础分类指标（precision, recall, F1-score, MCC）。
  - MCC = **0.4906**（中规中矩）。
- 训练模型二：**Logistic Regression**
  - 使用 `StandardScaler` 进行特征标准化。
  - 训练Logistic回归，验证集评估指标优于决策树：
    - Logistic F1 = **0.5140**
    - Logistic MCC = **0.5031**

2.3 模型选择与比较（Model Comparison）

- 比较两模型：
  - Logistic回归的召回率、F1得分、MCC均高于决策树。
- 决策：**选择 Logistic Regression 作为主力模型**，进入调优阶段。

2.4 Logistic回归模型调优（Logistic Regression Tuning）

- 使用 **GridSearchCV** 调整超参数：

- 搜索正则化强度 `C` 参数：[0.01, 0.1, 1, 10, 100]

其他参数也尝试调整了，但是耗时间太长，而且最优的是默认值，所以没必要加入model tuning

- 评估标准：**f1_score**
- 5折交叉验证。

- 结果：
  - 找到最佳 C。
  - 得到最佳交叉验证 F1-score。

2.5 （可删除）决策树调优（Decision Tree Tuning）

- 虽然逻辑回归更好，仍对决策树尝试GridSearch调优，防止意外：
  - `max_depth`，`min_samples_split`，`ccp_alpha`。
- 最终即便调优，决策树表现仍略差于Logistic Regression。

2.6 最佳模型保存（Model Saving）

- 将调优后的 Logistic Regression 最佳模型保存为：

```
./models/logistic_regression_best.pkl
```

Step 3: Task O2 - Age Prediction (Regression)

3.1 数据准备（Data Preparation）

- 选择特征集 `important_features_O2`：

```
["DIABETES", "COPD", "TOBACCO", "HYPERTENSION",
 "CARDIOVASCULAR", "RENAL_CHRONIC", "OBESITY",
 "ASTHMA", "INMSUPR"]
```

- 目标列：`AGE`
- 替换缺失值（97, 98, 99）为众数。
- 对特征进行标准化（StandardScaler），因为回归模型对特征尺度敏感。
- 划分训练集（80%）与验证集（20%）。

3.2 模型训练（Model Training）

训练了三种回归模型：

- **Random Forest Regressor**（100棵树，默认超参数）
- **Ridge Regression**（L2正则化）
- **Linear Regression**（普通线性回归）

（备注：**SVR** 因训练时间过长，需要7min，而且效果不如random forest，不是best，所以淘汰。）

3.3 模型评估（Model Evaluation）

使用以下指标评估每个模型：

- MAE（平均绝对误差）
- RMSE（均方根误差）
- R^2分数（决定系数）
- Pearson相关系数

各模型表现总结：

模型	MAE	RMSE	R^2 Score	Pearson Correlation
Random Forest	最低	最低	最高	最高
Ridge Regression	中等	中等	较低	较低
Linear Regression	最差	最差	最低	最低

✅ **Random Forest Regressor** 的综合表现最佳（最小误差、最高拟合度）。

3.4 可视化分析（Visualization）

- 绘制了 **真实值 vs 预测值散点图**（基于 Random Forest 模型）。

- 观察到大部分点接近理想对角线，但仍存在一定离散。

3.5 最佳模型保存（Model Saving）

- 保存最佳模型（Random Forest Regressor）到：

```
./models/random_forest_age_predictor.pkl
```

Step 4: Task O3 - Predict Age among Deceased Patients

4.1 数据准备（Data Preparation）

- 加载清洗过的数据集 `custom_covid19_cleaned.csv`。
- 筛选死亡患者（`DIED == 1`）。
- 特征选择：使用之前定义好的 `important_features_O3`：

```
["ICU", "INTUBED", "PNEUMONIA",
 "DIABETES", "HYPERTENSION", "OBESITY",
 "RENAL_CHRONIC", "COPD", "TOBACCO", "ASTHMA", "INMSUPR"]
```

- 目标列为 `AGE`。
- 替换缺失值（97,98,99）为众数。
- 特征标准化（StandardScaler）。
- 切分训练集与验证集（train 80% / val 20%）。

4.2 模型训练（Model Training）

- 使用 **Random Forest Regressor**（100棵树）训练模型。
- 没有进行更复杂的模型调优，因为本任务主要测试基础可行性。

4.3 模型评估（Model Evaluation）

评估指标：

- MAE（平均绝对误差）
- RMSE（均方根误差）
- R^2分数
- Pearson相关系数

结果观察：

- 相比之前的 Step 3，误差更大，R^2 和 Pearson 相关性明显下降。
- 说明**死亡患者的特征**（如重症、慢性病）与**年龄**之间的关系较弱，无法准确推断年龄。

4.4 保存最佳模型（Model Saving）

- 保存训练好的 Random Forest Regressor 到：

```
./models/random_forest_deceased_age_predictor.pkl
```

✅ 模型成功保存，即使效果一般，也作为完整流程的一部分保留。

Step 5: Task O4 - Feature Summary and Final Insights

5.1 重要特征总结（Top Features Summary）

任务 (Task)	重要特征 (Important Features)
O1（死亡预测）	AGE, ICU, INTUBED, PNEUMONIA, DIABETES, COPD, HYPERTENSION, OBESITY, RENAL_CHRONIC, CARDIOVASCULAR
O2（年龄预测）	DIABETES, COPD, TOBACCO, HYPERTENSION, CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, ASTHMA, INMSUPR
O3（死亡患者年龄预测）	ICU, INTUBED, PNEUMONIA, DIABETES, HYPERTENSION, OBESITY, RENAL_CHRONIC, COPD, TOBACCO, ASTHMA, INMSUPR

5.2 特征与任务关系总结（Attribute Importance Table）

任务 (Task)	简要观察 (Observations)
O1（死亡预测）	年龄与重症情况（如插管、ICU住院）是死亡风险的重要决定因素。
O2（年龄预测）	慢性病（如糖尿病、高血压）与年龄有一定相关性，但误差较大。
O3（死亡患者年龄预测）	单靠疾病严重程度特征，无法准确推断死亡患者的年龄。

5.3 项目总结（Final Observations）

- **O1（死亡预测）**

  死亡预测模型表现良好，逻辑回归比决策树略优。

  - 最佳模型：Logistic Regression（调参后）
  - F1得分和MCC表现可接受。

- **O2（年龄预测）**

  年龄预测存在一定误差，但整体趋势可以抓住。

  - 最佳模型：Random Forest Regressor
  - 误差比预期大，但属于数据本身性质导致。

- **O3（死亡患者年龄预测）**

  极具挑战性，死亡患者内部特征与年龄关系非常弱。

  - 最佳模型（形式上保存）：Random Forest Regressor
  - 但模型预测能力较弱，主要作为流程示范。

- **整体收获**

  通过本项目，完整实践了：

  - EDA
  - 特征工程
  - 模型训练与调优
  - 指标分析
  - 报告总结
  - 模型保存

完